United we stand: improving sentiment analysis by joining machine learning and rule based methods

Rentoumi, V ; Petrakis, S ; Klenner, M ; Vouros, G A ; Karkaletsis, V

United we stand: improving sentiment analysis by joining machine learning and rule based methods

Vassiliki Rentoumi*†, Stefanos Petrakis‡, Manfred Klenner‡, George A. Vouros †, Vangelis Karkaletsis*

*National Centre for Scientific Research "Demokritos", Inst. of Informatics and Telecommunications,
Athens Greece
vrentoumi,vangelis@demokritos.gr

†University of the Aegean, Dpt. of Information and Communication System Enginneering
Samos, Greece
georgev@aegean.gr

‡Institute of Computational Linguistics, University of Zurich,
Zurich, Switzerland
petrakis,klenner@cl.uzh.ch

Abstract

In the past, we have succesfully used machine learning approaches for sentiment analysis. In the course of those experiments, we observed that our machine learning method, although able to cope well with figurative language could not always reach a certain decision about the polarity orientation of sentences, yielding erroneous evaluations. We support the conjecture that these cases bearing mild figurativeness could be better handled by a rule-based system. These two systems, acting complementarily, could bridge the gap between machine learning and rule-based approaches. Experimental results using the corpus of the Affective Text Task of SemEval '07, provide evidence in favor of this direction.

1. Introduction

Exploiting figurative language using the machine learning system proposed in (Rentoumi et al., 2009) has opened an interesting new path to sentiment analysis. This specific machine learning method (FigML) has been trained on a corpus manually annotated with figurative language[1]. This method was more inclined towards the strong figurative use of language, like "Record-shattering day on Wall St". It has been observed, that the cases where the classification decision is taken within a small margin, are those bearing mild figurativeness, often yielding erroneous evaluations. Such a "marginal" case is:

(a) Ancient coin shows Cleopatra was no beauty.

In example (a), "beauty" extends its primary sense, and it is used as an expanded sense[2] denoting "a beautiful woman". Despite the use of this metaphorical extension, the negative polarity of this sentence can still be obtained without the need for word sense disambiguation. According to (Cruse, 2000), such cases of figurative language, are - synchronically - as literal as their primary sense, as a result of standardized usage.

Such cases are ideal candidates for a rule-based system like PolArt (Klenner et al., 2009) that has been designed to handle literal or semi-literal language through sentiment composition. PolArt combines the polarities of syntactic constituents like "beauty" and polarity shifters like "no" to compose the polarity of larger textual units.

We propose a method which aims at filling the gap for polarity detection in corpora where strong and mild figurative language are prominent. The proposed method, contrary to the one presented in (Rentoumi et al., 2009), which dealt with strong figurative language phenomena, deals more effectively with all expressions of figurative language, strong and mild. In this paper we introduce a novel method for sentiment analysis of figurative language, which overcomes the fallacies of the machine learning method attributed to the existence of mild figurative language, by delegating those to a rule-based approach (Klenner et al., 2009).

In particular this paper argues in favor of a collaborative approach to sentiment analysis consisting of two sub-methods acting complementarily: a machine learning method (Rentoumi et al., 2009) that handles the "non-marginal" cases which bear strong figurativeness and a compositional rule-based method (Klenner et al., 2009), for the "marginal" ones. Results verify that integrating a machine learning approach with a finer-grained linguistics-based one leads to a superior, best-of-breed coalition system.

In Section 2 we present related work and in Section 3 we describe our methodology. Our experimental setup and results are presented in Section 4. In Section 5 we present evaluation results. We conclude and list some ideas about future work in Section 6.

[1]Two subsets were extracted from the AffectiveText corpus (SemEval 07') and annotated with metaphors and expanded senses. They are available at: `http://www.iit.demokritos.gr/~vrentoumi/corpus.zip`

[2]expanded sense: a metaphorical extension/restriction of the word usage

2. Related Work

So far there is not much work in sentiment classification in sentences using supervised machine learning. The main reason recorded, is lack of sufficient labelled data for training. A recent work that describes sentiment classification of sentences, using statistical machine learning (SVM) is (Gamon and Aue, 2005). In (Andreevskaia and Bergler, 2008) the authors created a domain adaptive system for sentence and document level classification. Sentiment level classification of newspapers' headlines exploiting a variety of machine learning techniques has been the goal of the Affective Text Task of SemEval 07'(Strapparava and Mihalcea, 2007). A similar task, concerning the sentiment classification of headlines is addressed in (Rentoumi et al., 2009).

Moreover in (Rentoumi et al., 2009), structured models such as Hidden Markov Models (HMMs) are exploited in sentiment classification of headlines. In our collaborative approach presented here, the supervised machine learning method adopted (HMMs) in order to identify polarity in sentences bearing strong figurative language is identical to the one presented in (Rentoumi et al., 2009). The advantage of HMMs against other machine learning approaches employed till now in sentiment analysis is that the majority of them is based on flat bag-of-features representations of sentences, without capturing the structural nature of a sub-sentential interactions. On the contrary, HMMs being sequential models encode this structural information, since sentence elements are represented as sequential features.

On the other hand, rule-based approaches for sentiment detection in sentences are not used extensively. The rule-based approach employed in this paper originally presented in (Klenner et al., 2009) is based on the principle of compositionality. Moilanen and Pulman (Moilanen and Pulman, 2007) as well adopt a compositional approach to sentiment analysis.

We have also noticed that the bibliography for sentiment analysis is rather poor in methods which combine the benefits of machine learning together with rule-based approaches. In (Choi and Cardie, 2008) a combined method is presented performing expression level polarity classification which integrates inference-rules inspired by compositional semantics into learning. In this way the subsentential interactions are captured by compositional rules and learned by the system. The interaction of subsentential constituents, such as word level polarity and valence shifters yield the overall polarity of the whole phrase. Valence shifters originally studied in (Polanyi and Zaenen, 2004) are contextual elements in discourse that could modify the valence of opinionated words, thus affecting the overall polarity of a sentence. In our collaborative approach a manually annotated list of valence shifters is compiled and integrated in our system for sentence level polarity classification.

Our method for sentiment analysis combines in a complementary way two approaches, a machine learning together with a compositional rule-based one, and aims at revealing sentiment in figurative language.

3. Methodology Description

The proposed method involves four consecutive steps:

(a)Word sense disambiguation(WSD): We chose an algorithm[3] that assigns to every word in a sentence the WordNet sense that is mostly related to the WordNet senses of its neighbouring words, revealing the meaning of that word. This WSD algorithm takes as input each sentence of our corpus and a relatedness measure(Pedersen et al., 2005). The algorithm supports several WordNet based similarity measures, and among these, Gloss Vector (GV) (Pedersen et al., 2005) performs best for non-literal verbs and nouns (Rentoumi et al., 2008). GV exploits what is called second order co-occurance, which claims that two concepts (senses) are semantically related when they both occur with the same third concept. GV predicts the similarity for two WordNet senses by finding the cosine similarity of their respective Gloss Vectors. Integrating GV in the WSD step is detailed in (Rentoumi et al., 2009).

(b)Sense level polarity assignment(SLPA): We adopted a machine learning approach which exploits graphs based on character n-grams (Giannakopoulos et al., 2008). A (character) n-gram S^n contained in a text T can be any substring of length n of the original text. The *n-gram graph* is a graph $G = \{V^G, E^G, L, W\}$, where V^G is the set of vertices, E^G is the set of edges, L is a function assigning a label to each vertex and edge, and W is a function assigning a weight to every edge. n-grams label the vertices $v^G \in V^G$ of the graph. The (directed) edges are labeled by the concatenation of the labels of the vertices they connect in the direction of the connection. The edges $e^G \in E^G$ connecting the n-grams indicate proximity of these n-grams in the text within a given window D_{win} of the original text (Giannakopoulos et al., 2008). The edges are weighted by measuring the number of co-occurrences of the vertices' n-grams within the window D_{win}.

To compute models of polarity using n-gram graphs, we have used two sets of positive and negative examples of words and definitions provided by the General Inquirer [4] (GI). To represent a text set using n-gram graphs, we have implemented the *update/merge* operator between n-gram graphs of the same rank. Specifically, given two graphs, G_1 and G_2, each representing a subset of the set of texts, we create a single graph that represents the merging of the two text subsets: update$(G_1, G_2) \equiv G^u = (E^u, V^u, L, W^u)$, such that $E^u = E_1^G \cup E_2^G$, where E_1^G, E_2^G are the edge sets of G_1, G_2 correspondingly. The model construction process for each class (e.g. of the positive/negative polarity class) comprises the initialization of a graph with the first sense of a class, and the subsequent update of this initial graph with the graphs of the other senses in the class using the union operator. As we need the model of a class to hold the average weights of all the individual graphs contributing to this model, functioning as a representative graph for the class documents, the i-th graph that updates the class graph (model) uses a learning factor of $l = \frac{i-1}{i}, i > 1$.

The polarity class of each test sense, represented as n-gram

[3] http://www.d.umn.edu/~tpederse/senserelate.html

[4] http://www.wjh.harvard.edu/~inquirer/

graph exploiting its Synset and Gloss Example derived from WordNet, is determined by computing its similarity with the models of the classes: the class whose model is the most similar to the test sense n-gram graph, is the class of the document. The Graph Similarity exploited here in order to compare two graphs sets $\mathbb{G}_1, \mathbb{G}_2$ (one representing a sense and the other the model of a polarity class) is *Value Similarity (VS)* which indicates for every n-gram rank (Giannakopoulos et al., 2008), how many of the edges contained in graph G^i of rank n are also contained in graph G^j also of rank n, considering also the weights of the matching edges. VS is a measure converging to 1 for graphs that share their edges and have identical edge weights. Then the overall similarity VS^O of the sets $\mathbb{G}_1, \mathbb{G}_2$ is computed as the weighted sum of the VS over all ranks. More details for Graph Similarity are given in (Rentoumi et al., 2009).

c)HMMs training: HMMs serve two purposes, computing the threshold which divides the sentences in marginal/non-marginal and judging the polarity (positive/negative) of non-marginal sentences.

In the training step, two different HMMs are trained (one for the positive and one for the negative class) and the threshold for the distinction of marginal and non-marginal sentences is also computed. Only positive instances are used for the training procedure of HMMpos (the model for positive sentences) and only negative instances are used for training HMMneg (the model for negative sentences). After the extraction of the two models we use both positive and negative sentences to calculate the threshold for marginal/non-marginal cases. Each instance is tested with both trained models (positive and negative) and log probabilities are computed denoting the probability of the instance to belong to each model. For each polarity class we compute the absolute difference of the log probabilities and sort these differences in ascending order. The threshold for distinguishing marginal from non-marginal sentences is computed using the first Quartile (Q1) which separates the lower 25% of the sample population from the remaining data. Marginal cases are the ones below that threshold.

HMMs testing: Each instance in the testing data set is tested against both models (HMMpos and HMMneg) and the calculated log probabilities are used in order to decide if it is a marginal or a non-marginal case, according to the aforementioned threshold. If we decide that the instance is a non-marginal case the greater log probability (between HMMpos and HMMneg) provides us with the decision in which class (positive or negative) the specific instance belongs.

For our experiments we use data formed according to the format presented in (Rentoumi et al., 2009) and we perform ten fold cross validation approach for the evaluation step. For each fold 90% of the data are used for the training procedure and 10% for the testing step.

(d) Sentence-level polarity detection: The polarity of each sentence is determined by HMMs (Rentoumi et al., 2009) for non-marginal cases and by PolArt(Klenner et al., 2009) for marginal ones. PolArt employs compositional rules and obtains word-level polarities from a subjectivity lexicon (Wilson et al., 2005).

Example (a) from Introduction would be treated by PolArt in the following consecutive steps:

1. no$^{[DT:shifter]}$ beauty$^{[NN:positive]}$ $\rightarrow$ NEG$_1$ $^{[:negative]}$
2. was$^{[VBD:]}$ NEG$_1$$^{[:negative]}$ $\rightarrow$ NEG$_2$$^{[:negative]}$
3. Ancient$^{[JJ:]}$ coin$^{[NN:]}$ shows$^{[VVD:]}$ Cleopatra$^{[NP:]}$ NEG$_2$$^{[:negative]}$ $\rightarrow$ NEG$_3$$^{[:negative]}$

First, a determiner that operates as a polarity shifter is combined with a positive noun into NEG$_1$, a negative chunk. Then, a verb is combined with NEG$_1$ to produce NEG$_2$. Finally, the sentence's polarity is determined to be negative driven by NEG$_2$'s negative polarity.

The performance of FigML is added up with that of PolArt, and gives the total performance of the collaborative system.

4. Experimental Setup

4.1. Resources

We ran our experiments on three datasets:

- The AffectiveText corpus[5] from SemEval '07 comprising 1000 polarity annotated headlines (Strapparava and Mihalcea, 2007).

Figurative Language Datasets:

- 87 manually annotated headlines containing metaphors extracted from the AffectiveText corpus[6] (48neg/39pos) (Rentoumi et al., 2009).
- 190 manually annotated headlines containing expanded senses extracted from the AffectiveText corpus[5] (95neg/95pos) (Rentoumi et al., 2009).

We ran 4 variations of Polart, modifying the polarity lexicon it consults:

- SL: This is the subjectivity lexicon(Wilson et al., 2005). SL contains among others parts from the GI lexicon.
- SL+: This is the subjectivity lexicon(Wilson et al., 2005) with 54 added valence operators.
- Merged: The FigML system produces automatically 3 sense-level polarity lexica (AutSPs), one for each dataset. For the non-literal datasets (metaphors/expanded senses) these lexica target metaphors and expanded senses accordingly. For the AffectiveText dataset all word senses are targeted. 3 Merged lexica are produced by merging the SL+ lexicon with the AutSPs.
- MergedManual: We use 2 handcrafted sense-level polarity lexica (ManSPs)[7]. These lexica target metaphors and expanded senses accordingly. 2 MergedManual lexica are produced by merging SL+ with the ManSPs.

[5] http://www.cse.unt.edu/~rada/affectivetext/

[6] http://www.iit.demokritos.gr/~vrentoumi/corpus.zip

[7] ManSPs were produced by manually mapping expanded and metaphorical senses from Wordnet to GI (Rentoumi et al., 2009)

4.2. Experimental Results I (Figurative Language Datasets)

In the following sections experimental results are presented for both alternatives of the collaborative system, compared to the pure machine learning (FigML) method and the pure rule-based approach (Polart) for the Figurative Language Datasets, namely the datasets which are annotated with metaphorical expressions and expanded senses (see 4.1.).

4.2.1. Evaluation of the Systems on marginal cases

Focusing only on marginal cases we compared FigML with three variants of the rule-based method, Polart (using the SL lexicon), PolartSL+ (using the SL+ lexicon) and PolartMerged (using the Merged lexica) (see 4.1.). Table 1 presents performances in terms of recall (rec), precision (prec), f-score, for each polarity class (negatives/positives), for the four systems, across both figurative language datasets (metaphors (Met) and expanded senses (Exp)). All rule-based alternatives (Polart, PolartSL+, PolartMerged) outperform the pure machine learning (FigML) method. In particular PolartSL+ presents the best performance across all rule-based alternatives, for both data sets. Paired t-tests report that PolartSL+'s superior performance is statistically significant within: 2% statistical error threshold for both classes on the expanded senses dataset (p-value=0.02), while for metaphors' data set we cannot support within 5% statistical error that PolartSL+ performs better than FigML (p-value=0.54).

		Polart		PolartSL+		PolartMerged		FigML	
		neg	pos	neg	pos	neg	pos	neg	pos
Met	rec	0.687	0.785	0.875	0.785	0.687	0.642	0.562	0.714
	prec	0.785	0.687	0.823	0.846	0.687	0.642	0.692	0.588
	fscore	0.733	0.733	0.848	0.814	0.687	0.642	0.620	0.645
Exp	rec	0.652	0.644	0.529	0.783	0.411	0.594	0.235	0.594
	prec	0.625	0.667	0.529	0.783	0.318	0.687	0.210	0.628
	fscore	0.638	0.653	0.666	0.740	0.511	0.603	0.292	0.472

Table 1: *Performance scores for marginal cases on Metaphors and Expanded senses data sets*

4.2.2. Evaluation of the Systems on non marginal cases

We have claimed that the non-marginal cases are the ones that bear strong figurativeness, and can be better treated by the machine learning method (FigML) which is originally designed to treat such cases.

Focusing on non-marginal cases we compared FigML, with three variants of the rule-based method, Polart (using the SL lexicon), PolartSL+ (using the SL+ lexicon) and PolartMerged (using the merged lexica). Table 2 presents performances in terms of recall (rec), precision (prec), f-score, for each polarity class for the four systems across both figurative language data sets (metaphors (Met) and expanded senses (Exp) datasets). FigML outperforms all rule-based (Polart) variants. Such an observation strengthens our initial conjecture that FigML can effectively treat strong figurative language.

		Polart		PolartSL+		PolartMerged		FigML	
		neg	pos	neg	pos	neg	pos	neg	pos
Met	rec	0.687	0.800	0.593	0.840	0.812	0.720	0.937	0.640
	prec	0.814	0.667	0.826	0.617	0.787	0.750	0.769	0.888
	fscore	0.745	0.727	0.690	0.711	0.800	0.734	0.845	0.744
Exp	rec	0.694	0.685	0.652	0.757	0.694	0.585	0.777	0.728
	prec	0.694	0.685	0.734	0.679	0.632	0.650	0.746	0.761
	fscore	0.694	0.685	0.691	0.716	0.662	0.616	0.761	0.744

Table 2: *Performance scores for non-marginal cases on Metaphors and Expanded senses data sets*

4.2.3. Evaluation of the Systems on full data sets

For the complete datasets we compared the pure machine learning method (FigML) and the pure rule-based method (Polart) with two variants of the collaborative system, CollabSL+ (using the SL+ lexicon) and CollabMerged (using the Merged lexica). Table 3 presents scores for each polarity class, across both figurative language datasets (metaphors (Met) and expanded senses (Exp)).

For the majority of cases, both system variants outperform FigML[8] and Polart. This fact leads to the conclusion that a collaborative system which combines the virtues of a rule-based and a machine learning method, can more properly handle a corpus which bristles with strong and mild figurative language, than two separate language specific systems functioning alone. Another observation that can be made from Table 3 and Table 2, is that FigML's performance drops for the full data sets (Table 3) while pure Polart's performance remains the same both for the non-marginal data set (Table 2) and the full data sets (Table 3). According to the aforementioned observation we can rationally attribute FigML's performance drop to the existence of marginal cases, a fact that was also our primary intuition.

Paired t-tests report that CollabSL+'s superior performance compared to FigML's is statistically significant within 2% statistical error threshold for both polarity classes on the expanded senses data set (p-value = 0.019). On the other hand, for the metaphors data set we cannot support within 5% statistical error that CollabSL+ is better than FigML's (p-value = 0.13).

		Polart		CollabSL+		CollabMerged		FigML	
		neg	pos	neg	pos	neg	pos	neg	pos
Met	rec	0.687	0.794	0.916	0.693	0.854	0.641	0.812	0.666
	prec	0.804	0.673	0.785	0.871	0.745	0.781	0.750	0.742
	fscore	0.741	0.729	0.846	0.771	0.796	0.704	0.780	0.702
Exp	rec	0.684	0.673	0.736	0.747	0.705	0.705	0.652	0.673
	prec	0.677	0.680	0.744	0.739	0.705	0.705	0.666	0.659
	fscore	0.680	0.677	0.740	0.743	0.705	0.705	0.659	0.666

Table 3: *Performance scores for full system runs on Metaphors and Expanded senses data sets*

Moreover, we cannot support within 5% statistical error that CollabSL+ is better than Polart for the metaphors (p-value = 0.13) or the expanded senses data set (p-value = 0.10).

4.3. Experimental results II (whole data set)

In the following sections experimental results are presented for both alternatives of the collaborative system, compared to the FigML method for the whole data set (1000 headlines).

[8]The FigML system presented here is almost identical to the one presented in (Rentoumi et al., 2009). The activation of valence shifters, for the experiments on the AffectiveText dataset, results in an increased performance.

These experiments were further performed in order to have a comparison with FigML under a more extended data set, although none of these methods, neither the collaborative nor the FigML system were originally designed to handle raw (unannotated corpus), which in its majority consists of literal language.

4.3.1. Evaluation of the systems on marginal cases

In this experiment, focusing on marginal cases we compared FigML with two variants of Polart, PolartSL+ (using the SL+ lexicon) and PolartMerged using the Merged Lexica. It is important to mention that only one third of the Affective Text corpus contains figurative language (metaphors and expanded senses), whereas the remaining sentences (700) are considered literal. Therefore we have an additional reason to claim that most marginal cases would probably belong to literal language, which is why a rule-based system such as Polart was selected to treat them. Table 4 presents scores for each polarity class (negative/positive), for the three systems, concerning the whole AffectiveText corpus (All) (see section 4.1.). For this extended data set of marginal cases - the corpus is bigger, thus marginal cases are relatively more - our initial intuition is getting verified, since both rule-based alternatives (PolartSL+, PolartMerged), perform better than pure machine learning FigML. In particular paired t-tests report that PolartSL+'s superior performance is statistical significant within 1% statistical error threshold across both polarity classes on the Affective Text data set (p-value = 0.011).

		PolartSL+		PolartMerged		FigML	
		neg	pos	neg	pos	neg	pos
All	rec	0.609	0.725	0.492	0.572	0.546	0.516
	prec	0.696	0.642	0.543	0.522	0.538	0.524
	fscore	0.650	0.681	0.516	0.546	0.542	0.520

Table 4: *Performance scores for marginal cases on the Affective Text corpus*

4.3.2. Evaluation of the systems on full data set

For the complete data set of the Affective Text corpus we compared the performance of FigML (already presented in (Rentoumi et al., 2009)) with two variants of the collaborative system, CollabSL+ and CollabMerged. Table 5 presents scores for each polarity class concerning the Affective Text corpus (All). Both alternatives of the collaborative system (CollabSL+, CollabMerged) outperform FigML. This fact can leads us to the conclusion that performance boost obtained with the use of the rule-based approach propagates to the overall performance of the system. Moreover our proposed approach performed consistently well for an extended data set. It also performs well even in mixed corpora, where style varies, and figurative language coexists with literal language.

		CollabSL+		CollabMerged		FigML	
		neg	pos	neg	pos	neg	pos
All	rec	0.612	0.604	0.583	0.560	0.597	0.545
	prec	0.638	0.577	0.603	0.542	0.601	0.543
	fscore	0.624	0.588	0.593	0.551	0.599	0.544

Table 5: *Performance scores for full system runs on the Affective Text corpus*

Paired t-tests report that CollabSL+'s superior performance compared to FigML's is statistically significant within: 1% statistical error threshold across both polarity classes on the AffectiveText dataset (p-value= 0.011).

5. Evaluation of the Collaborative Method

For thoroughly evaluating the proposed collaborative method we need to test our basic working assumption: Compositional rules work sanely, so that Polart's fallacies are a result of erroneous polarities upon which the rules are applied.

To test this, we will run the CollabMergedManual system, a variation of the CollabMerged system, this time exploiting the MergedManual lexica (see 4.1.). Doing so, we shall assess the role of the ManSPs lexica within the MergedManual lexica as a performance boost factor of the CollabMergedManual system.

Table 6 presents the performance of the CollabMergedManual against that of CollabSL+ and FigML for both polarity classes, tested upon the metaphors (Met) and expanded senses (Exp) datasets. Note that FigML is also exploiting the ManSPs lexica (see 4.1.).

		CollabSL+		CollabMergedManual		FigML	
		neg	pos	neg	pos	neg	pos
Met	rec	0.770	0.795	0.791	0.820	0.750	0.846
	prec	0.822	0.738	0.844	0.761	0.857	0.733
	fscore	0.795	0.765	0.817	0.790	0.800	0.786
Exp	rec	0.778	0.863	0.789	0.842	0.757	0.873
	prec	0.850	0.796	0.833	0.800	0.857	0.783
	fscore	0.813	0.828	0.810	0.820	0.804	0.825

Table 6: *Manual Evaluation of the Collaborative system*

Concerning the expanded senses dataset we did not observe any performance boost of CollabMergedManual relative to CollabSL+. For the CollabMergedManual system, 41% of the word polarities were contributed by the ManSPs lexica, which includes polarized words introduced by the human experts that either did not exist in the SL+ lexicon or that existed in the SL+ but the experts assigned them with a different polarity than the one in SL+. Another 30% was found in both the ManSPs and the SL+ lexicon, which we can call the agreement percentage. And finally, the remaining 29% was found only in the SL+ lexicon.

Concerning the metaphors dataset we do observe, in both positive and negative classes, a performance boost of CollabMergedManual relative to CollabSL+. For the CollabMergedManual system, 57% of the word polarities were contributed by the ManSPs lexica, which includes polarized words introduced by the human experts that either did not exist in the SL+ lexicon or that existed in the SL+ but the experts assigned them with a different polarity than the one in SL+. Another 26% was found in both the ManSPs and the SL+ lexicon, which we can call the agreement percentage. And finally, the remaining 17% was found only in the SL+ lexicon.

Reflecting on these measurements we can see that a higher degree of participation of the ManSPs lexica for the metaphors dataset ($57\% > 41\%$) leads to an noticeable performance boost. We can also see that for expanded senses the agreement between manually disambiguated lexica and the Subjectivity lexicon is slightly higher($30\% > 26\%$) than

for the metaphors dataset, which could point to a more literal reading for the expanded senses dataset. Despite the small size of our sample, we conjecture that our collaborative system has potential to further performance gains if we integrate a fully manual sense-level polarity lexicon for all words of both datasets. That remains to be tested with a broader manually prepared polarity lexicon.

6. Conclusions and Future Work

We present a novel collaborative methodology for sentiment analysis. It provides evidence for the complementarity of a rule-based compositional system (PolArt) and a machine learning system (FigML). Sentiment composition proved its applicability as PolArt treated successfully the marginal cases that would otherwise cause FigML's performance to drop. Experiments showed that the performance boost for marginal cases gets propagated to the overall performance of the collaborative system surpassing the machine learning approach.

The initial observation that marginal cases bear mild figurativeness and are therefore treated by PolArt effectively is supported by the experimental results.

We will test the collaborative method on a more extensive corpus. Since correct sense-level polarity is vital for the evolution of the collaborative system, we intend to dynamically produce proper sense-level polarity lexica exploiting additional machine learning approaches (e.g. SVMs).

7. References

A. Andreevskaia and S. Bergler. 2008. When specialists and generalists work together: overcoming domain dependence in sentiment tagging. In *Proceedings of ACL-08: HLT, pages:290–298, 2008*.

Y. Choi and C Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, 793–801, 2008, Association for Computational Linguistics*.

D.A. Cruse. 2000. *Meaning in language*. Oxford University Press.

M. Gamon and A. Aue. 2005. Automatic identification of sentiment vocabulary exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 57–64*.

G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3).

M. Klenner, S. Petrakis, and A. Fahrni. 2009. Robust compositional polarity classification. In *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007)*, pages 378–382, Borovets, Bulgaria, September 27-29.

T. Pedersen, S. Banerjee, and S. Patwardhan. 2005. Maximizing semantic relatedness to perform word sense disambiguation. *Supercomputing institute research report umsi*, 25.

L. Polanyi and A. Zaenen. 2004. Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–9.

V. Rentoumi, V. Karkaletsis, G. Vouros, and A. Mozer. 2008. Sentiment analysis exploring metaphorical and idiomatic senses: A word sense disambiguation approach. *Proccedings of International Workshop on Computational Aspects of Affectual and Emotional Interaction (CAFFEi 2008)*.

V. Rentoumi, G. Giannakopoulos, V. Karkaletsis, and G. Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

C. Strapparava and R. Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the thInternational Workshop on Semantic Evaluations (SemEval 2007), Prague, Czeck Republic*, pages 70–74.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*.